Paola Merulla - Angiola Pollastri[1]

Improvement of the estimator of the mean in presence of partial respondents

1. Introduction

Even in the best survey, despite all efforts (motivating and interesting introductions, well-designed and not too long questionnaires, solicited participation, and so on), some data will be missing.
In fact, in many sample surveys, some of the units selected do not provide an acceptable measure of the variable of interest, denoted by Y. We shall indicate as non-respondent those units. These missing values not only mean less efficient estimates because of the reduced size of the data base but also that the standard complete-data methods cannot be immediately used to analyse the data.
Moreover, possible biases exist because the respondents are often systematically different from the non-respondents. These biases are difficult to eliminate since the precise reasons from nonresponse are usually not known. For instance, in a mail survey the questionnaire may not be returned or returned but not completely filled in. In a survey with personal interviews, some individuals refuse to respond to some or all the questions . It may occur also interviewer errors: omitted questions, illegible recording of the response and so on.

Here we shall consider the situation in which for some units the principal item concerning the variable Y is missing but we have the response to another variable, called X.

Two general techniques have been proposed in literature to face the above problem:
1) imputation technique (see, for instance, Little and Rubin (1987), Trimarchi (1990)). This operation considers a prevision of the missing value. Imputation implies that an imputed value $\hat{y}_i$ is produced for a missing value y_i, that is, $\hat{y}_i$ fills the blank for the missing value in the data analysis
2) adjustments of the Horvitz- Thompson estimator through the estimated individuals response probabilities in order to compensate for the non-respondents (see Giommi (1984), Yogendra P. Chaubey et al (1997)).

In the present paper we want to improve the imputation technique when we have information about an auxiliary variable prior to sampling or during the operation of sampling.

We will considerate two estimators generally used in practise. The first one is based only on the complete responses and it is generally biased. The second one considers the auxiliary information available at sample level.

[1] This work was discussed together by the two authors. In particular, §2-3 were carried out by Pollastri and §4 by Merulla. Address for communications: Pollastri Angiola – Dipartimento di Metodi Quantitativi per l'Economia – Università degli Studi di Milano – Bicocca – P.za Ateneo Nuovo, 1 – 20126 – Milano – Italy. E-mail: Angiola.Pollastri@unimib.it – pmerull@tin.it

Considering that it may happen to have also auxiliary information at population level, here we propose a new estimator which is an application of the regression technique. We shall show that the new estimator is more efficient than the estimators considered before.
The above demonstration is done considering the situation that the number of non respondents to item regarding the variable Y is known.
Then we analize through simulation studies the properties and the distribution of the estimators considered in different situations

2. Estimation

Let consider a finite population P of N elementary units denoted by $u_1,....,u_i,...,u_N$. Associated with each member of P there is a value of the variable Y indicated by y_i^* and an auxiliary variable X which assumes value x_i^*. Let assume that the variable X may be the predictor of the variable *Y.*
Let suppose that the relation between Y and X be

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad i\text{=1,...,N}$$

where $E(\varepsilon_i) = 0$ and ε_i and X_i are independent r.v. and $E(\varepsilon_i) = 0$.
Here we shall suppose that β is known.

It follows that

$$E(Y_i) = \alpha + \beta\mu_x$$

Let suppose that ideally the population may be divided in two groups: in the first the units will respond to the item connected with the variable Y and the related mean will be indicated by μ_{Y_1}; in the second group they will not respond and the mean of the variable Y will be denoted by μ_{Y_2}. Suppose that *X* will receive responses from all the units. We denote by μ_{X_1} and μ_{X_2} respectively the mean of X in the group where Y obtains responses and in the group where Y will be missing.

Let suppose to select a simple random sample of size n.
We denote by n_1 the number of complete observations and by $n_2 = n - n_1$ the number of missing data on variable *Y.* Before selecting the sample, n_1 is a random variable because we do not know how many units will respond to Y. After having collected the data, n_1 will be a constant.

2.1 Estimation without using auxiliary information

If we consider only the complete responses, indicated by $y_{1i}\,(i = 1,..., n_1)$, we estimate μ_y by

$$\bar{y}_1 = \frac{\sum_{i=1}^{n_1} y_{1i}}{n_1}$$

The estimator $\overline{Y}_1$ is affected by the following bias

$$\delta_{\overline{Y}_1} = E(\overline{Y_1}) - \mu_y = W_2(\mu_{Y1} - \mu_{Y2})$$

where W_2 is the proportion of non-respondents in the population.
The demonstration is analogous to the situation of the total non-respondents (see Pollastri, 1997).

The variance of the r.v. $\overline{Y}_1$ is

$$Var(\overline{Y_1}) = \frac{\sigma_y^2}{n_1}$$

The mean square error is given by

$$MSE(\overline{Y_1}) = \frac{\sigma_y^2}{n_1} + W_2^2(\mu_{Y_1} - \mu_{Y_2})^2$$

2.2 Estimation using information from the sample: Regression Imputation

The estimation of μ_y (see Trimarchi, 1990) may be

$$\hat{\mu}, = \bar{y}_1 + \beta(\bar{x} - \bar{x}_1)$$

Where $\bar{x}_1$ is the mean of the variable *X* computed in the part of the sample where the observations are complete. $\bar{x}$ is the mean of the observations concerning the variable X in all the sample.
It is interesting to observe that μ_{yi}, corresponds to

$$\hat{\mu}_y = r[\frac{1}{n_1}\sum_1^{n_1} y_{1i}] + (1-r)[\frac{1}{n_2}\sum_1^{n_2} \hat{y}_{2i}]$$

that is, $\hat{\mu}_y$ is the mean of the sampled values y_{1i} in the stratum of the respondents and of the imputed values

$$\hat{y}_{2i} = \bar{y}_1 + \beta(x_{2i} - \bar{x}_1)$$

where y_i is missing. The pounds are constituted by the rate of response in the sample $r = \frac{n_1}{n}$ and the rate of non-response $(1-r) = \frac{n_2}{n}$.

The estimator $\hat{\mu}_y$ is unbiased.
In fact

$$\begin{aligned}
&E(\hat{\mu}_y) - \mu_Y = E(\overline{Y}_1) - \mu_Y + \beta E(\overline{X} - \overline{X}_1) = \\
&W_2(\mu_{Y_1} - \mu_{Y_2}) + \beta\mu_X - \beta\mu_{X_1} = \\
&W_2(\alpha + \beta\mu_{X_1} - \alpha - \beta\mu_{X_2}) + \beta\mu_X - \beta\mu_{X1} = \\
&W_2\beta(\mu_{X_1} - \mu_{X_2}) - W_2\beta(\mu_{X_2} - \mu_{X_1}) = 0
\end{aligned}$$

The variance of the estimator $\hat{\mu}_y$ is given by

$$\begin{aligned} Var(\hat{\mu}_y) &= E(\hat{\mu}_y - \mu_y)^2 \\ &= E[(\overline{Y}_1 - \mu_y)^2 + \beta^2(\overline{X} - \overline{X}_1)^2 + 2\beta(\overline{X} - \overline{X}_1)(\overline{Y}_1 - \mu_y)] \end{aligned}.$$

Adding and subtracting μ_x in the second term, we can obtain

$$\begin{aligned} &Var(\hat{\mu}_y) \\ &= Var(\overline{Y}_1) + \beta^2 E(\overline{X} - \mu_X)^2 + \beta^2 E(X_1 - \mu_X)^2 + 2\beta^2 E[(\overline{X} - \mu_X)(\dddot{X}_1 - \mu_y)] \\ &+ 2\beta E[\overline{X}(\overline{Y}_1 - \mu_y)] - 2\beta E[\overline{X}_1(\overline{Y}_1 - \mu_y)] \end{aligned}$$

Remembering (Pollastri, 1997) that

$$E[(\overline{X} - \mu_x)(\overline{Y} - \mu_y)] = \frac{\sigma_{xy}}{n}$$

and that

$$\overline{X} = \frac{\overline{x}_1 n_1 + \overline{X}_2 n_2}{n}$$

and observing that

$$E[(\overline{X}_2 - \mu_x)(\overline{X}_1 - \mu_x)] = 0,$$

after some algebra, we arrive to the result

$$Var(\hat{\mu}_y) = \frac{\sigma_y^2}{n_1}(1 - \frac{n_2}{n}\rho^2)$$

All the above considerations are done when the data are collected, that is when n_1 and n_2 are fixed.

2.3 Imputation using auxiliary information from population.

Let suppose we know the real average μ_x of the variable X in the population. For instance, if P is constituted by the population of a city, X may represent the age or if the population is constituted by all the farms of a region, X may be the area of each farm. In this situation we can consider the estimator

$$\hat{\mu}_Y^{'} = \overline{y}_1 + \beta(\mu_x - \overline{x}_1)$$

This corresponds to impute the missing data with

$$\hat{y}_{2i} = \overline{y}_{lr} + \beta(x_{2i} - \overline{x}_1) = \overline{y}_1 + \beta(x_{2i} - \mu_x)$$

The estimator $\hat{\mu}'_y$ is unbiased. In fact

$$\begin{aligned} E(\hat{\mu}'_y) - \mu_Y &= E(\overline{Y}_1) - \mu_Y + \beta E(\mu_x - \overline{X}_1) \\ &= W_2(\mu_{Y_1} - \mu_{Y_2}) + \beta\mu_X - \beta\mu_{X_1} \\ &= W_2\beta(\mu_{X_1} - \mu_{X_2}) - W_2\beta(\mu_{X_1} - \mu_{X_2}) = 0 \end{aligned}$$

The variance of the estimator $\hat{\mu}'_y$ is given by

3. Comparison between estimators

The efficiency of the estimator $\hat{\mu}'_y$ with respect to the estimator $\hat{\mu}_y$ is given by

$$E_{\hat{\mu}'_y/\hat{\mu}_y} = \frac{1-\rho^2}{1-\rho^2\,\dfrac{n_2}{n}}$$

In Table 1 are reported the values of the efficiency as a function of the correlation coefficient and of the rate of non-response.
From these values it is evident that the efficiency of the estimator $\hat{\mu}'_y$, which considers also the knowledge of μ_x, increases when the rate of non-response decreases, keeping fixed ρ.
Furthermore the efficiency of the estimator $\hat{\mu}'_y$ with respect to the estimator $\hat{\mu}_y$ is greater when the correlation coefficient is very high.

Table 1

n_2/n ρ,	0.2	0.5	0.8
0.10	0.964	0.769	0.384
0.30	0.972	0.811	0.445
0.50	0.98	0.857	0.529

4. Simulation purpose

Now we aim to examine missing data problems[2] and to explain the main characteristics of the different approaches presented in this paper.
The purpose is to analyse the properties of different estimators under specific situations; in particular we wish to observe:

1. The estimators behaviour on varying the number of cases in the selected sample and the number of respondents.

[2] This has been made by simulations created using FORTRAN 90 software.

2. The eventual variation of expected value and variance of the estimator due to the estimation of β when it is unknown.
3. The estimators characteristics when missing data are:
    - 3.1. Completely at random (MCAR)
    - 3.2. High values of the variable Y
    - 3.3. Low values of the variable Y
    - 3.4. Values of Y on high values of the variable X (MAR)
    - 3.5. Values of Y on low values of the variable X (MAR)

It will be finally interesting to observe the estimators distributions when the population considered is normally and when it is lognormally distributed. In the following table are reported the symbols used.

Table of symbols

V	preceding an estimator mark denotes its variance and with the addition of a final S it represents the estimated variance
E	preceding an estimator mark, denote the expected value
MSE	preceding an estimator mark, denote the mean square error
MCAR	Completely at random (MCAR)
OR	High values of the variable Y
LY	Low values of the variable Y
HXMAR	Values of Y on high values of the variable X (MAR)
LXMAR	Values of Y on low values of the variable X (MAR)
BE	after mark estimator means that β is estimated with sample data

4.1. Simulation steps

For the simulation it was created an artificial population with known mean and variance.
After having selected a random sample from the population, we have remarked the observations (x_i, y_i) respectively value of the auxiliary variable X and of the variable to study Y.
To simplify the results interpretation we have assigned value 0 to the mean of Y in the population and 0.64 to its variance. We have also chosen a high correlation coefficient (0.9) between X and Y as we wished to describe an example of strong relation.
Once the sample has been drawn out, we have proceeded with the missing data selection of the variable Y. As mentioned before we have considered different missing data mechanism:

- Missing completely at random (MCAR): missingness is unrelated to the values of the variables
- Missing at random (MAR): given survey variables, the missing data distribution depends only on variables that are entirely recorded in the data set
- Missing data which depend on the same variable on which we enter nonrespondents

As we shall illustrate in the description of simulation results, the presence of MCAR allows complete case analysis without danger of biased estimation, though some loss in efficiency.
As MCAR is often difficult to meet in practice, it is useful to include in the survey variables which have low probability to record nonrespondents and that are highly correlated with the variable to study. Actually it has been previously demonstrated that, with imputation of missing data or with

the application of regression with auxiliary population information, ML estimators can improve the estimation, as we see in the next pages. For each missing data mechanism exposed we have calculated the mean estimation in every sample extracted with the following estimators:

a) Complete Case Estimator ($\overline{Y}_1$)
b) Regression Imputation Estimator ($\hat{\mu}_y$)
c) Regression Imputation Estimator with auxiliary information from the population ($\hat{\mu}'_y$)

4.2. Simulation results

4.2.1 *First analysis: expected values and variances examination*

Table 2

Estimators expected values – *s*=1000

n=100 n_2=20	$E(\hat{\beta})$	$E(\overline{Y}_1)$	$E(\hat{\mu}_y)$	$E(\hat{\mu}'_y)$
MCAR		-0,00178	-0,00040	0,00089
MCAR_BE	1,43775		-0,00030	0,00112
HY		-0,28001	-0,05465	-0,05336
HY_BE	1,26795	-0,28001	-0,08125	-0,0801
LY		0,27639	0,05194	0,05323
LY_BE	1,26721	0,27639	0,07853	0,07997

On examining the table 1, we immediately note that the estimators appear to be unbiased for random missing data. This table has been created for different sample sizes and numbers. Obviously we have indicated only one kind of table but we are able to affirm that there is a general improvement on extending the sample size and the number of samples extracted, but for the high variability assumed, this improvement is not regular.
Among the considered estimators, $\hat{\mu}'_y$ seems to come nearer to the population mean of Y than the other estimators.
Let now consider the missing data that are not completely at random. If they depend on the variable Y we note significant changes. First the $\overline{Y}_1$ estimator is not unbiased yet while $\hat{\mu}'_y$ and $\hat{\mu}_y$ estimators are in general unbiased, especially for large samples even if they present worse values than MCAR case.

As regards missing data depending on Y the mean estimations are fairly biased.
Let observe now the variances of the estimators.

Table 3

Estimated variance – *s*=1000

n=100 n_2=20	$\hat{V}(\bar{Y}_1)$	$\hat{V}(\hat{\mu}_y)$	$\hat{V}(\hat{\mu}'_y)$
MCAR	0,00806	0,00646	0,00150
MCAR_BE	0,00806	0,00647	0,00154
HY	0,00697	0,00683	0,00152
HY_BE	0,00697	0,00686	0,00162
LY	0,00667	0,00642	0,00145
LY_BE	0,00667	0,00677	0,0017

Y mean and variance and estimators variance calculated with theoretical formulisations

	$E(\bar{Y})$	$V(\bar{Y})$	$V(\bar{Y}_1)$	$V(\hat{\mu}_y)$	$V(\hat{\mu}'_y)$
n=100, n_2=20	-0,00168	0,00620	0,00800	0,00670	0,00152

We can immediately note that $\hat{\mu}'_y$ variance is even lower than mean estimator variance without nonrespondents. This occurs because of the missing data imputation, which causes data homogenisation and consequently a reduction of the variance. Nevertheless it is important to stress that $\hat{\mu}'_y$ is the unbiased estimator with lowest variance.

Moreover, from the simulations carried on we have observed that the more the sample size raises the more the variance decreases. This trend happens even in the event of non-random missing, where we notice not particular differences with the same variance estimators in MCAR.

4.3.2 β estimated through sample data

When missing data are completely at random β estimation is unbiased; only in the event of small sample size (ex. *s*=50, *n*=50 and *m*=10) *b* is fairly far from the true value of β.

Table 4

β Estimation: *s*=50, missing data at random

n=50 n_2=10	$E(\hat{\beta})$
MCAR_BE	1,40345
HY_BE	1,26522
LY_BE	1,25584

In case of non-random missing data *b* is biased; the bias decreases with the sample size raising. The effect of β estimation on mean estimators expected value is not particularly relevant for MCAR;

even for the other missing data mechanisms there are minimal differences which vanish when *N* increases, though their expected value are a little worse than the firsts.

4.2.2 The effect of nonrespondents fraction increase

Table 5

Random missing data – *s*=500

	$E(\overline{Y}_1)$	$E(\hat{\mu}_y)$	$E(\hat{\mu}'_y)$	$\hat{V}(\overline{Y}_1)$	$\hat{V}(\hat{\mu}_y)$	$\hat{V}(\hat{\mu}'_y)$
n=500 n_2=50	-0,00826	-0,00737	-0,00198	0,00143	0,00131	0,00029
n=500 n_2=100	-0,00653	-0,00687	-0,00148	0,00145	0,00129	0,00031
n=500 n_2=150	-0,00700	-0,00796	-0,00257	0,00168	0,00141	0,00035

	$MSE(\overline{Y}_1)$	$MSE(\hat{\mu}_y)$	$MSE(\hat{\mu}'_y)$
n=500 n_2=50	0,0015	0,00136	0,00029
n=500 n_2=100	0,0015	0,00133	0,00031
n=500 n_2=150	0,00173	0,00148	0,00036

Table 6

Missing data depending on Y: High values of Y – *s*=500

	$E(\overline{Y}_1)$	$E(\hat{\mu}_y)$	$E(\hat{\mu}'_y)$	$\hat{V}(\overline{Y}_1)$	$\hat{V}(\hat{\mu}_y)$	$\hat{V}(\hat{\mu}'_y)$
n=500 n_2=50	-0,28179	-0,05708	-0,04918	0,00136	0,00130	0,00028
n=500 n_2=100	-0,28615	-0,05964	-0,05425	0,00145	0,00136	0,00032
n=500 n_2=150	-0,40331	-0,08162	-0,07623	0,00156	0,00141	0,00036

$MSE(\overline{Y}_1)$	$MSE(\hat{\mu}_y)$	$MSE(\hat{\mu}'_y)$
0,02773	0,0026	0,00122
0,08333	0,00491	0,00326
0,16422	0,00807	0,00617

Table 7

Missing data depending on Y: Low values of Y – *s*=500

	$E(\overline{Y}_1)$	$E(\hat{\mu}_y)$	$E(\hat{\mu}'_y)$	$\hat{V}(\overline{Y}_1)$	$\hat{V}(\hat{\mu}_y)$	$\hat{V}(\hat{\mu}'_y)$
n=500 n_2=50	0,14828	0,02266	0,02805	0,00130	0,00127	0,00028
n=500 n_2=100	0,27223	0,04600	0,05139	0,00137	0,00126	0,00029
n=500 n_2=150	0,38948	0,06803	0,07342	0,00145	0,00128	0,00034

	$MSE(\overline{Y}_1)$	$MSE(\hat{\mu}_y)$	$MSE(\hat{\mu}'_y)$
n=500 n_2=50	0,02328	0,00178	0,00106
n=500 n_2=100	0,07548	0,00338	0,00293
n=500 n_2=150	0,15315	0,00591	0,00573

On examining the above tables, when the missing data fraction arises, it is confirmed what has been demonstrated in the conceptual part of this paper, such as:

- On augmenting the proportion of nonrespondents there is a worsening of the estimators expected value in case of missing data depending on Y;
- There is a general variance growth.
- Nonrespondents fraction extremely affects $\hat{\mu}_y$ estimator on determining its variance, owing to the particular variance function form.

Let consider the variance of estimator $\hat{\mu}_y$.We can clearly note that the value it assumes depends not only on the population variance and correlation coefficient which furthermore represent determinant components, but as well on nonrespondents number. As regards $\hat{\mu}'_y$, however, the more this number increases the more its variance grows, owing to the reduction of both denominator and weight preceding the correlation coefficient, though it preserves a smaller variance value than $\overline{Y}_1$ and $\hat{\mu}_y$.

4.3 Estimators' distributions when the distribution of the variables is normal

Once verified the quality, i.e. efficiency of $\hat{\mu}'_y$ estimator, it is now useful to analyse the studied estimators' distribution. This could represent a further property especially for normal distributions (for example, let think about the possibilities achieved from hypothesis tests and confidence intervals). To obtain the real distribution, we should consider all samples *s*, which could be extracted according to a specific sample design. For each *a (a: 1...s)* we should know both the probability *p(a)* to extract *s* and the population mean estimator value $\hat{y} = \hat{y}(s)$. It would be possible, then, to calculate the real expected value, bias and variance of $\hat{y}$. However this is generally an impossible task, as the potential samples number is very high. This is the reason for using simulations to study estimators' statistical properties in sample surveys.

In the following charts and tables, consequently, we shall pay attention to the estimators' distribution form. In the charts we will illustrate only $\hat{\mu}'_y$ distribution. We immediately note that all estimators are normal distributed, thus preserving the population distribution from which the sample has been extracted.

Chart 1. Estimators distribution with random missing data

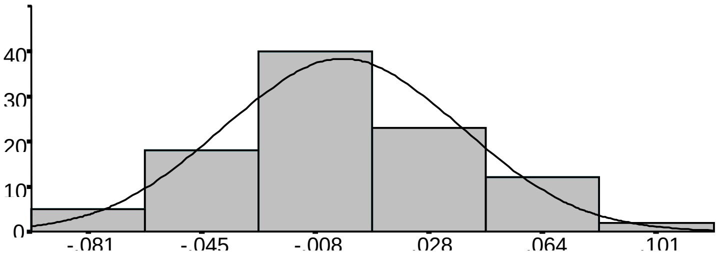


Chart 2. Estimators distribution with missing data depending on X

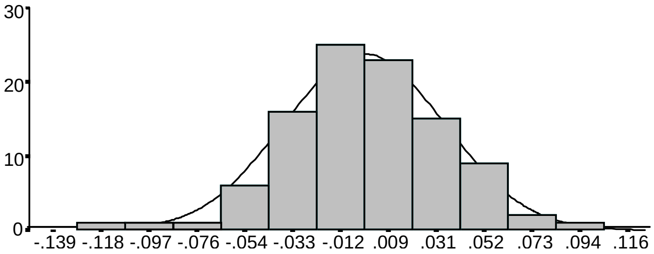
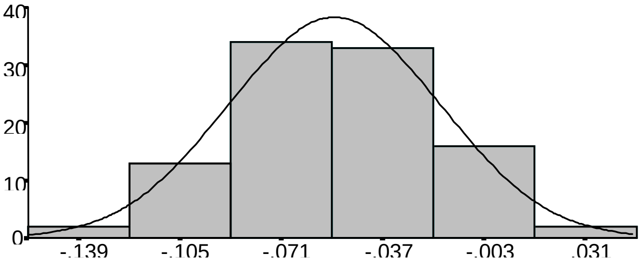


Chart 3. Estimators distribution with missing data depending on X



Note: Estimator = $\hat{\mu}'_y$ *, s=100, n=100,* n_2*=20*

Table 7 instead, illustrates a test used in Monte Carlo simulation to calculate *the actual confidence level, which* may be used to study different estimators comparisons.

Let consider fix population and sample design. A high number of samples is drawn out from the population under the considered sample design. Once extracted, the sample is placed again in the population before the next sample selection, in order to draw out the sample always from the same population. The samples number is high and it is indicated with *K*. If *K* is sufficiently high the distribution of the *K* estimations (empirical sample distribution) is able to approximate truly enough the real sample distribution, that is not easily obtainable, as we explained before.

We can therefore calculate

$$\bar{\hat{y}} = \frac{1}{K}\sum_{j=1}^{K}\hat{y}_j$$

which is an expected value estimation $E(\hat{y})$

$$S^2_{\hat{y}^2} = \frac{1}{K-1}\sum_{j=1}^{K}\left(\hat{y}_j - \bar{\hat{y}}\right)^2$$

which is a variance estimation $V(\hat{y})$ and finally,

$$\bar{\hat{V}} = \frac{1}{K}\sum_{j=1}^{K}\hat{V}(\hat{y})$$

which is a variance expected value estimation, E[$\hat{V}(\hat{y})$].

If for each sample we calculate the 95% approximate level confidence interval

$$\hat{y} \pm 1{,}96\left[\hat{V}(\hat{t})\right]^{1/2}$$

and we successively count the intervals number R, which contain the true value μ_y, then R/K is the actual confidence level estimation. The actual confidence level estimation could differ from 95% as $\left(\bar{\hat{y}} - \mu_y\right)\Big/\left[\hat{V}\left(\bar{\hat{y}}\right)\right]^{1/2}$ has only an approximate normal distribution.

In the following table it is illustrated a comparison between $\hat{\mu}'_y$ and $\bar{Y}_1$.

Table 8
The actual confidence level. *s*=100, *n*=100, *ns*=20

	$\hat{\mu}'_y$		$\bar{Y}_1$	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.9$
MCAR	0.94	0.94	0.94	0.92
MCAR_BE	0.97	0.95		
HY	0.13	0.63	0.05	0.06
HY_BE	0.08	0.47		
LY	0.25	0.76	0.06	0.11
LY_BE	0.16	0.61		
HXMAR	0.98	0.95	0.60	0.11
HXMAR_BE	0.93	0.94		
LXMAR	0.95	0.94	0.57	0.20
LXMAR_BE	0.96	0.91		

Firstly, we note that for high values of Y as missing data (HY and HY_BE), the confidence level is extremely low, that means that the number of intervals containing the real μ_y is very small; the estimators considered are therefore unreliable in those cases. We wish to stress that the number of intervals is created with estimators values and variances.
The same matter takes place in the opposite situation, where minimal values of Y are considered as missing: though the presence of a modest improvement compared to the previous case, the actual confidence level is lower than 95%.

These results are therefore useful to argue that the expected value of analysed estimators particularly affects their reliability even if their variance has regularly presented favourable results in accordance with the theory.
A further consideration regards the different results according to the correlation coefficient ρ. As we could expect , the $\hat{\mu}'_y$ estimator results greatly improved in case of missing data depending on Y, though they remain under 95%. This demonstrates what theory explains: the opportunity of using auxiliary information, and in particular in this case the knowledge of the true mean μ_y, produces an estimation improvement especially when X and Y are strongly related.
When we oppositely consider the situations where X values determine missing data, $\hat{\mu}'_y$ estimator is extremely satisfactory, conforming therefore a normal distribution.
If we analyse the results paying attention on the differences between $\hat{\mu}'_y$ and $\bar{Y}_1$, we can also argue other interesting considerations.
Firstly, we note that, in case of random missing data, the differences between these estimators are minimal and moreover both are satisfactory. However, when ρ increases there is a little improvement of $\hat{\mu}'_y$ compared to $\bar{Y}_1$ for the reasons previously explained.
As regards the other situations, we observe moreover that $\bar{Y}_1$ presents a very low actual confidence level. Thus it results particularly worst even for cases where $\hat{\mu}'_y$ don't seem to be extremely adequate. We should think about the minimal probabilities of finding the true parameter of interest in the intervals created with $\bar{Y}_1$ estimator and on the possible biased interpretations that such estimator could erroneously involve.

4.4 Comparison with a *lognormal* population distribution

As announced early in the paper, the simulation tested for samples extracted from population normally distributed, has then been conducted for population with a *lognormal* distribution.
The aim of this operation was to verify if in presence of a different distribution or, in worst cases, of non-knowledge of the population distribution, the estimator found could improve in any case the mean estimation. Moreover, there would be an important result if it was normally distributed as the previous case.
In the following tables we wish to illustrate the main results (we have limited the exposition to the case of s=100)

Table 9

Estimators expected values – *s*=100

n=100 n$_2$=20	$E(\hat{\beta})$	$E(\overline{Y}_1)$	$E(\hat{\mu}_y)$	$E(\hat{\mu}'_y)$
MCAR		1,36366	1,37295	1,37465
MCAR_BE	1,35603	1,36366	1,37252	1,37398
HY		1,37709	1,3699	1,37161
HY_BE	1,20858	1,37709	1,37034	1,36913
LY		1,62776	1,41916	1,42086
LY_BE	1,59854	1,62776	1,39384	1,38907

Note: $\overline{Y}_l$ *=1,37713*

Estimated variance – *s*=100

n=100 n$_2$=20	$\hat{V}(\overline{Y}_1)$	$\hat{V}(\hat{\mu}_y)$	$\hat{V}(\hat{\mu}'_y)$
MCAR	0,02386	0,02121	0,00671
MCAR_BE	0,02386	0,02131	0,00724
HY	0,02511	0,02094	0,01032
HY_BE	0,02511	0,02123	0,00805
LY	0,0293	0,02367	0,00787
LY_BE	0,0293	0,01635	0,00427

Y mean and variance and estimators variance calculated with theoretical formulisations

	$E(\overline{Y})$	$V(\overline{Y})$	$V(\overline{Y}_1)$	$V(\hat{\mu}_y)$	$V(\hat{\mu}'_y)$	$V(\overline{Y}_l)$
n=100 n$_2$=20	1.37154	0.02082	0.02125	0.01881	0.00903	1.7006

In the table above exposed, we note that as the normal distribution case, when missing are at random, the estimators considered are unbiased. However, on comparing with the previous situation, we immediately note that when missing data depend on Y, $\hat{\mu}'_y$ is no more the best estimator, as $\hat{\mu}_y$ and even $\hat{\mu}'_y$ seem sometimes register better values. As regards variance, let pay attention to the high variance of the mean estimator ($V(\overline{Y}_l)$) which can evidently produce some biased result.

4.4.1 The effect of nonrespondents fraction increase

As the simulation first part, we have then analysed the estimators characteristics on increasing the number of missing data. In this case we have added even tables on MAR mechanisms to explore the complete field. We can observe that there are not great differences when nonrespondents increase, excluding the situations where missing data depend on X. Here we particularly note that $\bar{Y}_1$ estimator is strongly biased on comparison with the other estimators. Finally, we can denote that the YMP and $\hat{\mu}'_y$ estimators have similar trend and values in all cases illustrated.

Table 10
Random missing data – *s*=100

	$E(\bar{Y}_1)$	$E(\hat{\mu}_y)$	$E(\hat{\mu}'_y)$	$\hat{V}(\bar{Y}_1)$	$\hat{V}(\hat{\mu}_y)$	$\hat{V}(\hat{\mu}'_y)$
n=500 n_2=50	1,37237	1,37422	1,37503	0,00327	0,00307	0,00083
n=500 n_2=100	1.37942	1.37710	1.37790	0,00386	0,00315	0,00102

Missing data depending on Y: High values of Y – *s*=100

	$E(\bar{Y}_1)$	$E(\hat{\mu}_y)$	$E(\hat{\mu}'_y)$	$\hat{V}(\bar{Y}_1)$	$\hat{V}(\hat{\mu}_y)$	$\hat{V}(\hat{\mu}'_y)$
n=500 n_2=50	1,37572	1,3771	1,37791	0,00343	0,0032	0,00086
n=500 n_2=100	1.37303	1.37535	1.37615	0,00377	0,00323	0,00091

Missing data depending on Y: Low values of Y – *s*=100

	$E(\bar{Y}_1)$	$E(\hat{\mu}_y)$	$E(\hat{\mu}'_y)$	$\hat{V}(\bar{Y}_1)$	$\hat{V}(\hat{\mu}_y)$	$\hat{V}(\hat{\mu}'_y)$
n=500 n_2=50	1,50113	1,39571	1,39651	0,00369	0,00329	0,00093
n=500 n_2=100	1.63466	1.42658	1.42738	0,00452	0,0037	0,00116

Missing data depending on X: High values of X – *s*=100

	$E(\bar{Y}_1)$	$E(\hat{\mu}_y)$	$E(\hat{\mu}'_y)$	$\hat{V}(\bar{Y}_1)$	$\hat{V}(\hat{\mu}_y)$	$\hat{V}(\hat{\mu}'_y)$
n=500 n_2=50	1,08762	1,30192	1,30272	0,00171	0,00221	0,00046
n=500 n_2=100	0.94098	1.28231	1.28311	0,00135	0,00185	0,00039

Missing data depending on Y: Low values of X – *s*=100

	$E(\bar{Y}_1)$	$E(\hat{\mu}_y)$	$E(\hat{\mu}'_y)$	$\hat{V}(\bar{Y}_1)$	$\hat{V}(\hat{\mu}_y)$	$\hat{V}(\hat{\mu}'_y)$
n=500 n_2=50	1,4951	1,38181	1,38261	0,00368	0,00331	0,00093
n=500 n_2=100	1.61972	1.39475	1.39556	0,00453	0,00367	0,00117

In this next table there is represented the actual confidence level, calculated only for a correlation coefficient of 0.9. It seems that when missing data depend on high values of Y or low values of X both actual confidence values are weak, while in the other situations $\hat{\mu}'_y$ presents best results as in case of normal distributed population.

Table 11
The actual confidence level. *s*=100, *n*=100, *ns*=20

$\rho = 0.9$	$\hat{\mu}'_y$	$\overline{Y}_1$
MCAR	0,96	0,93
MCAR_BE	0,96	
HY	0,93	0,95
HY_BE	0,94	
LY	0,94	0,68
LY_BE	0,95	
HXMAR	0,37	0
HXMAR_BE	0,2	
LXMAR	0,94	0,73
LXMAR_BE	0,85	

It is, finally interesting to examine graphically the forms of $\hat{\mu}'_y$ estimator distribution which seems confirm the previous results.

Note: Estimator = $\hat{\mu}'_y$*, s=100, n=100, ns=20 – population lognormally distributed*

Chart 1. Estimators distribution with random missing data

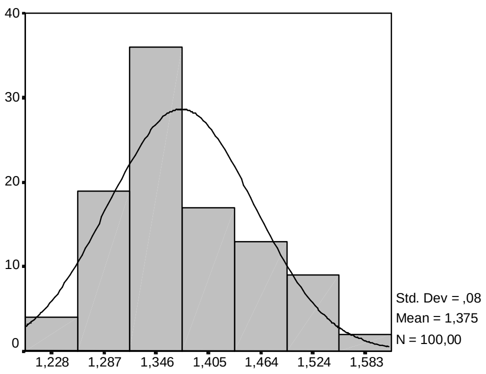


Chart 2. Estimators distribution with missing data depending on X

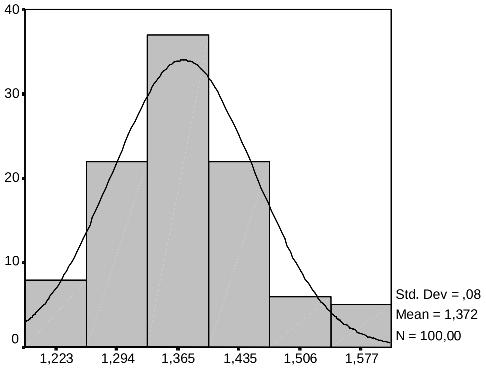
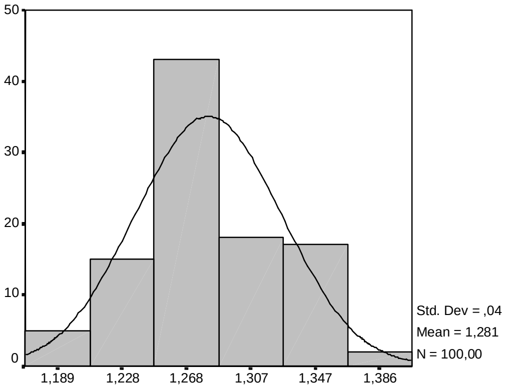


Chart 3. Estimators distribution with missing data depending on X



5. Conclusions

In this paper we have proposed an estimator of the mean useful when some information regarding the variable to be studied are missing but we know all the observations of an auxiliary variable in the sample and the mean of this variable in the population. Using this information we can improve the estimator used in the above situation. The variance of the new estimator is computed and so it is possible to study the efficiency.
The simulation study confirms the properties of the estimators. Moreover it is possible to conclude that the distribution of the considered estimators is normal when the distribution of the variables X and Y is normal in the population. This property is not preserved when the variables X and Y are lognormal in the population.

References

1) AA.VV (1994), Fortran. Subroutines for statistical applications, Vol. 1-2, IMSL-STAT/LIBRARY.
2) Brainerd W.S., Goldberg C. H., Adams J. C. (1994), Programmer's Guide to Fortran 90, 2nd ed., UNICOMP, New Mexico.

3) Cicchitelli G, Herzel A., Montanari G.E. (1992), Il campionamento statistico, Il Mulino, Bologna.

4) Cochran W.C. (1977), Sampling Tecniques, J. Wiley and Sons, New York, third edition.

5) Fadini B., Savy C. (1991), *F 77 S* . Il Fortran strutturato, Vol. 1, CUEN, Napoli.

6) Giommi A.(1984), Sulla stima delle probabilità di risposta nel campionamento da popolazioni finite, Atti della XXXII Riunione Scientifica della Società Italiana di Statistica, Sorrento.

7) Kalton, G. e Kaprzyk D. (1986), The treatment of Missing Survey Data, Survey Methodology, Vol. 12 n.1, pp.1-16.

8) Kalton, G. e Kaprzyk D.(1982), Imputing for missing survey response, Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 146-151.

9) Lessler J. T., Kalsbeek W. D. (1992), Nonsampling Error in Surveys, J. Wiley & Sons, New York.

10) Nordholt E.S. (1998), Imputation: Methods, Simulation, Experiments and Practical Examples, International Statistical Review, 66, pp. 157-180.

11) Pollastri A. (1997), Elementi di Teoria dei Campioni, CUSL, Milano.

12) Pollastri A. (1999), Stimatori col metodo della regressione in caso di non rispondenti con un campionamento doppio, Rapporto di ricerca del Dipartimento di Metodi Quantitativi per l'Economia,n.4, Università degli studi di Milano-Bicocca.

13) Rossi P.H., Wright J.D., Andersen B. (ed.) (1983), Handbook of survey research, Academic Press.

14) Rubin D.B. (1987), Multiple Imputation for Nonresponse in Surveys, J. Wiley & Sons, New York.

15) Sarndal, C.E., Svensson, B. e Wretman, J.(1992), Model assisted survey sampling, Springer Verlag, New York.

16) Trimarchi F. (1990), L’ imputazione dei dati mancanti nelle indagini campionarie: una applicazione delle tecniche di regressione, Banca d’ Italia-Temi di discussione.

17) Yogendra P. Chaubey, Crisalli A.N.(1997), Adjustment of the inclusion probabilities in the case of nonresponse, Quaderni di Statistica e Matematica Applicata alle Scienze Economico-Sociali, Vol.XIX, n.1-2, pp.53-65.